# Supplementary Material for SDFormer: Vision-based 3D Semantic Scene Completion via SAM-assisted Dual-channel Voxel Transformer

Yujie Xue, Huilong Pi, Jiapeng Zhang, Yunchuan Qin, Zhuo Tang, Kenli Li, Ruihui Li*
College of Computer Science and Electronic Engineering, Hunan University
{xueyj, phl880217, zhangjp, qinyunchuan, ztang, lkl, liruihui}@hnu.edu.cn

## A. Overview

In this supplementary material, we first provide more quantitative and qualitative analyses of our proposed SDFormer, along with some experimental comparison results. Next, we include additional implementation details, such as descriptions of the metrics and details of the training process. Finally, we will discuss the limitations of our work and outline directions for future research in the last section.

## B. Additional results

### B.1. More Quantitative Results

In this section, we present quantitative results of our method on the SemanticKITTI validation set in Table 1, comparing it with the state-of-the-art camera-based methods for a more comprehensive evaluation. Compared to other baselines, our method achieves significant improvements in mIoU, demonstrating its effectiveness in semantic scene completion. Additionally, the significant improvement in mIoU also demonstrates the superiority of our constructed semantic volume. Specifically, our method shows notable improvements in capturing structured objects (e.g., roads, sidewalks, buildings) and some small objects (e.g., poles and traffic lights).

### B.2. More Qualitative Results

We report more qualitative results in Figure 1. It can be observed that, compared to other camera-based methods, our approach shows significant improvements, especially in cluttered scene layouts and occluded areas. As seen in the third row and the sixth and seventh rows of the figure, our results are closer to the ground truth at intersections. Additionally, in terms of long distances, both quantitative and qualitative analyses show that SDFormer predicts more complete and accurate results. For example, the cars in the first and second rows and the distant trees have clearer segmented outlines.

*Corresponding authors.

### B.3. More Experiment Results

#### B.3.1. Effect of the SAM model type

The Table 2 shows that using larger SAM model types can enhance overall performance, but it also significantly increases the model's parameter count. Additionally, it demonstrates that SAM's performance can affect the overall model performance. Moving forward, we will focus on further optimizing the use of Vision Foundation Models.

#### B.3.2. Ablation study for Semantic Calibration Affinity

In the Table 3, we conduct additional ablation experiments on the Semantic Calibration Affinity (SCA) module to validate our choice of layer configuration. We see that while different layer counts yield comparable IoU values, the mIoU differs. Using two transformer layers achieves a good balance in model performance.

#### B.3.3. Ablation study for Cross Attention fusion schemes

In the Table 4, we ablated the cross-attention mechanism for the final two voxel fusions, using simpler addition and multiplication fusion methods. This resulted in a significant drop in mIoU, indicating that the two voxels still experience information misalignment. Our adaptive cross-attention mechanism better integrates features from the two different types of voxels.

## C. Implementation Details

### C.1. Metrics

Following [8], the main consideration of SSC is the mean Intersection over Union (mIoU), which considers the IoU of all semantic classes for prediction without considering the free space. The mIoU is calculated by:

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TN_c + FP_c + FN_c} \tag{1}$$

where $TP_c$, $TN_c$, $FP_c$, and $FN_c$ are the true positives, ture nagatives, false positives and false negatives predic-

| Methods | IoU | road (15.30%) | sidewalk (11.13%) | parking (1.12%) | other-grnd (0.56%) | building (14.10%) | car (3.92%) | truck (0.16%) | bicycle (0.03%) | motocycle (0.03%) | other-vehicle (0.20%) | vegetation (39.3%) | trunk (0.51%) | terrain (9.17%) | person (0.07%) | bicylist (0.07%) | motorcyclist (0.05%) | fence (3.90%) | pole (0.29%) | traf.-sign (0.08%) | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LMSCNet[7]† | 28.61 | 40.68 | 18.22 | 4.38 | 0.00 | 10.31 | 18.33 | 0.00 | 0.00 | 0.00 | 0.00 | 13.66 | 0.02 | 20.54 | 0.00 | 0.00 | 0.00 | 1.21 | 0.00 | 0.00 | 6.70 |
| AICNet[5]† | 29.59 | 43.55 | 20.55 | 11.97 | 0.07 | 12.94 | 14.71 | 4.53 | 0.00 | 0.00 | 0.00 | 15.37 | 2.90 | 28.71 | 0.00 | 0.00 | 0.00 | 2.52 | 0.06 | 0.00 | 8.31 |
| JS3C-Net[12]† | 38.98 | 50.49 | 23.74 | 11.94 | 0.07 | 15.03 | 24.65 | 4.41 | 0.00 | 0.00 | 6.15 | 18.11 | 4.33 | 26.86 | 0.67 | 0.27 | 0.00 | 3.94 | 3.77 | 1.45 | 10.31 |
| MonoScene[1] | 37.12 | 57.47 | 27.05 | 15.72 | 0.87 | 14.24 | 23.55 | 7.83 | 0.20 | 0.77 | 3.59 | 18.12 | 2.57 | 30.76 | 1.79 | 1.03 | 0.00 | 6.39 | 4.11 | 2.48 | 11.50 |
| TPVFormer[3] | 35.61 | 56.50 | 25.87 | 20.60 | 0.85 | 13.88 | 23.81 | 8.08 | 0.36 | 0.05 | 4.35 | 16.92 | 2.26 | 30.38 | 0.51 | 0.89 | 0.00 | 5.94 | 3.14 | 1.52 | 11.36 |
| OccFormer[13] | 36.50 | 58.85 | 26.88 | 19.61 | 0.31 | 14.40 | 25.09 | **25.53** | 0.81 | 1.19 | 8.52 | 19.63 | 3.93 | 32.62 | 2.78 | 2.82 | 0.00 | 5.61 | 4.26 | 2.86 | 13.46 |
| VoxFormer-S[6] | 44.02 | 54.76 | 26.35 | 15.50 | 0.70 | 17.65 | 25.79 | 5.63 | 0.59 | 0.51 | 3.77 | 24.39 | 5.08 | 29.96 | 1.78 | 3.32 | 0.00 | 7.64 | 7.11 | 4.18 | 12.35 |
| VoxFormer-T[6] | 44.15 | 53.57 | 26.52 | 19.69 | 0.42 | 19.54 | 26.54 | 7.26 | 1.28 | 0.56 | 7.81 | 26.10 | 6.10 | 33.06 | 1.93 | 1.97 | 0.00 | 7.31 | 9.15 | 4.94 | 13.35 |
| HASSC-T[9] | 44.58 | 57.23 | 29.08 | 19.89 | **1.26** | 20.19 | 27.33 | 17.06 | 1.07 | 1.14 | 8.83 | 27.01 | 7.71 | 33.95 | 2.25 | **4.09** | 0.00 | 7.95 | 9.20 | 4.81 | 14.74 |
| Symphonies[4] | 41.92 | 56.37 | 27.58 | 15.28 | 0.95 | 21.64 | 28.68 | 20.44 | 2.54 | 2.82 | **13.89** | 25.72 | 6.60 | 30.87 | **3.52** | 2.24 | 0.00 | 8.40 | 9.57 | 5.76 | 14.89 |
| H2GFormer-T[10] | 44.69 | 57.00 | 29.37 | 21.74 | 0.34 | 20.51 | 14.29 | 6.80 | 0.95 | 0.91 | 9.32 | **27.44** | 7.80 | 36.26 | 1.15 | 0.10 | 0.00 | 7.98 | 9.88 | 5.81 | 14.29 |
| **SDFormer (ours)** | **45.65** | **64.67** | **33.20** | **22.16** | 0.02 | **25.35** | **33.75** | 20.09 | **3.25** | **3.13** | 9.56 | 26.93 | **9.07** | **38.99** | 2.94 | 2.72 | 0.00 | **11.42** | **11.98** | **7.27** | **17.18** |

Table 1. Quantitative results on the SemanticKITTI validation set. † represents the results obtained when these methods use RGB inputs, which are implemented and reported in MonoScene [1]. The best results are in **Bold**.

| **SDFormer Setting** | IoU(%)↑ | mIoU(%)↑ | #Param |
|---|---|---|---|
| + ViT-L | 45.96 | 17.36 | 308M |
| + ViT-B | 45.65 | 17.18 | 91M |

Table 2. Ablation study for SAM model type.

| **SCA Setting** TransFormer Layer | IoU(%)↑ | mIoU(%)↑ |
|---|---|---|
| 0 | 44.81 | 16.66 |
| 1 | **45.87** | 16.65 |
| 2 | 45.65 | **17.18** |
| 3 | 45.55 | 16.55 |

Table 3. Comparison with the different number of transformer layers in the SCA module on model performance on the SemanticKITTI validation set.

| **Setting** | IoU(%)↑ | mIoU(%)↑ |
|---|---|---|
| Add-Conv Fusion | 45.72 | 16.74 |
| Multiplicative Fusion | 45.90 | 16.60 |
| Cross Attention (Ours) | 45.65 | 17.18 |

Table 4. Ablation study for other fusion schemes on the SemanticKITTI validation set.

tions for class $c$. Note the strong interaction between IoU and mIoU, as better geometric estimation (i.e., high IoU) can be achieved by invalidating semantic labels (i.e., low mIoU).

| Methods | Inference Mem.(M) | Inference Times(s) | FLOPs | IoU(%) | mIoU(%) |
|---|---|---|---|---|---|
| TPVFormer | 6391 | 0.295 | 1031G | 34.25 | 11.26 |
| OccFormer | 7454 | 0.258 | 969G | 36.50 | 13.46 |
| BRGScene | 6172 | 0.315 | 1518G | 43.85 | 15.43 |
| HTCL-S | 10408 | 0.357 | 1924G | 45.51 | 17.13 |
| SDFormer(Ours) | 8994 | 0.376 | 1748G | 45.65 | 17.18 |

Table 5. Efficiency analysis on SemanticKITTI val set.

## C.2. Semantic Constructor More details

According to the method in [11], when we input the image $I_l$ and its segmentation feature map $I_{sam}$, we treat $I_{sam}$ as an augmented version of $I_l$ and experiment with $I_{sam}$ for supervision. Following [2], for each pixel we optimize the loss for the best matching source image use a form of self-supervision,

$$L_p = \min_n pe(I_l, I_{sam}) \tag{2}$$

where $pe$ as a combination of SSIM and $L_1$ losses. Next, we define a set of ordered planes $P$, each perpendicular to the optical axis, linearly spaced between depths $d_{min}$ and $d_{max}$. The image is encoded into a depth feature map $F_d$, which is distorted to its viewpoint using each assumed alternative depth $d \in P$ and the already estimated relative positions. This generates a distorted feature map $F_{warp}$, and the absolute difference between features distorted at each depth and those from $I_l$ constructs the final cost volume.

## D. Limitation and Future Works

As shown in Table 5, our method slightly increases the burden, which brings corresponding performance improvement, which is acceptable. Even compared with the latest sota HTCL-S, using multi-frame input, SDFormer outperforms it in both memory consumption and performance. Although SDFormer demonstrates strong performance in benchmarks, the model's inference speed can still be improved. Since SSC is primarily used in autonomous driving or unmanned systems, a lightweight network would be suitable for various hardware for deployment. Therefore, we take model lightweight as the future work to make the model more conducive to downstream applications.

## References


[1] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 2

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2

[3] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 2

[4] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20258–20267, 2024. 2

[5] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020. 2

[6] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 2

[7] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 2

[8] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 1

[9] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14792–14801, 2024. 2

[10] Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5722–5730, 2024. 2

[11] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1164–1174, 2021. 2

[12] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3101–3109, 2021. 2

[13] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 2

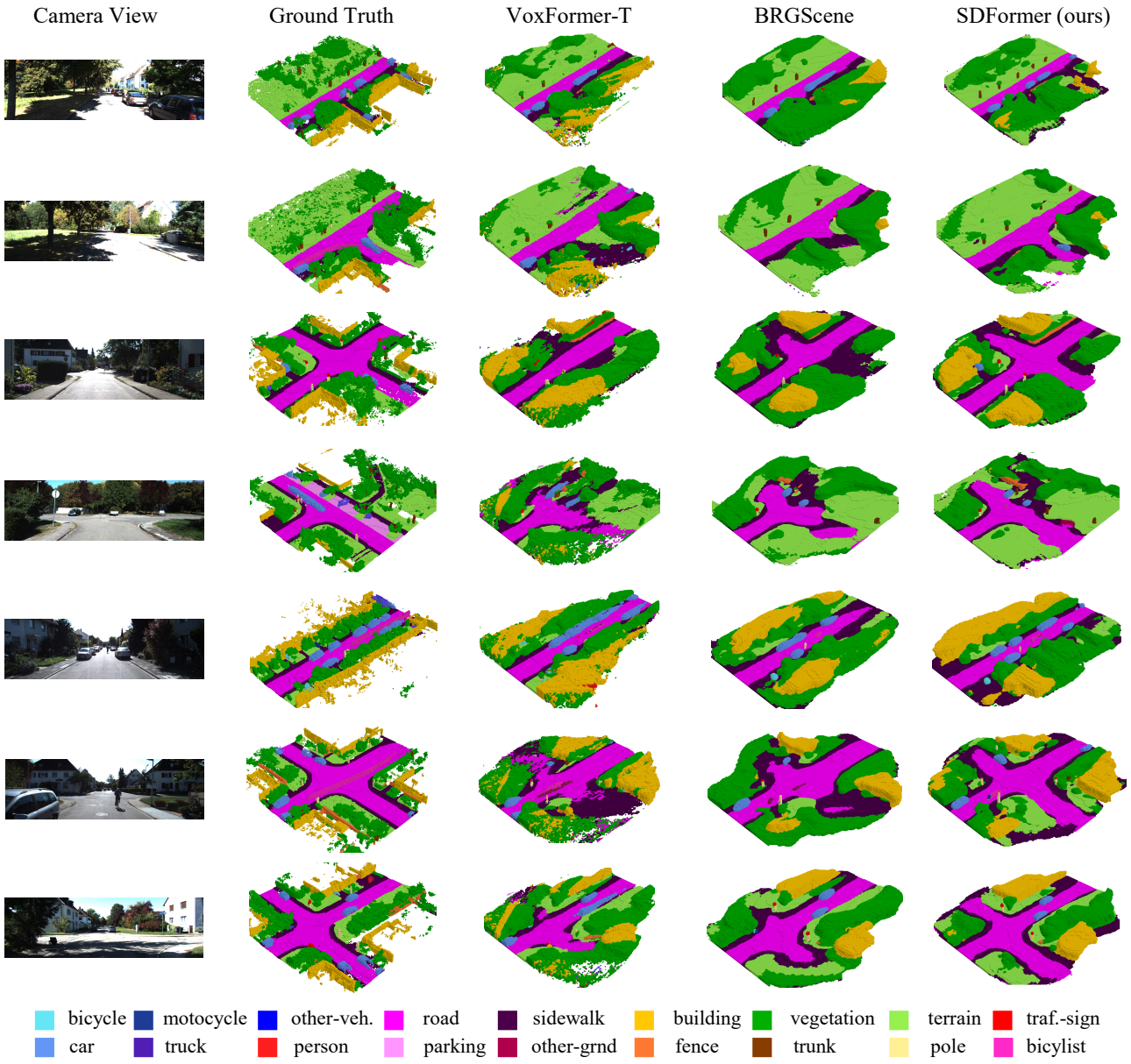


Figure 1. The visual results of our SDFormer and the state-of-the-art methods on the SemanticKITTI validation set. Please zoom in for details.